DOCUMENT RESUME

ED 464 117 TM 033 798

AUTHOR Yang, Wen-Ling; Dorans, Neil J.; Tateneni, Krishna
TITLE Sample Selection Effect on AP Multiple-Choice Score to Composite Score Scaling.
PUB DATE 2002-04-00
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cutting Scores; *Equated Scores; Multiple Choice Tests; Sample Size; *Sampling; Selection
IDENTIFIERS *Advanced Placement Examinations (CEEB); *Composite Scores; Equipercentile Equating; Free Response Test Items; Linear Equating Method; Linking Metrics; Smoothing Methods

ABSTRACT

Scores on the multiple-choice sections of alternate forms are equated through anchor-test equating for the Advanced Placement Program (AP) examinations. There is no linkage of free-response sections since different free-response items are given yearly. However, the free-response and multiple-choice sections are combined to produce a composite. Therefore, to derive new-form cut scores on the composite score scale that are comparable to the old-form cut scores, the multiple-choice score of the AP examinations are linked to the composite score through single-group linking design. This study investigates whether the multiple-choice to composite linking functions remain invariant over subgroups by region for two AP examinations using 3 years of test data. The region groups are of interest because the AP program administers different free-response sections to different time-zone regions as a precaution for security reasons. The study focused on: (1) how invariant cut scores are across regions; and (2) whether the small sample size for some regional groups presents particular problems for assessing linking invariance. Both equipercentile and linear linking methods were applied. The equatability index proposed by N. Dorans and P. Holland (2000) is used to evaluate the invariance of the linking functions, and the cross-classification approach is used to evaluate the invariance of the composite cut scores. Overall, the curvilinear linkings across regions seem to hold up reasonably well. Nevertheless, more examinations are needed to decide whether the linkings in the small regions contain enough data to support the use of weak models such as equipercentile scaling. The strong smoothing associated with the linear linking model, as expected, reduces the variability of linkages across regions. However, more work is needed to see if less strong models can be found for improving consistency over the unsmoothed linkings without incurring the bias apparent with the linear linkings. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 117

Sample Selection Effect on AP Multiple-choice Score to Composite Score Scaling

Wen-Ling Yang, Neil J. Dorans and Krishna Tateneni

Educational Testing Service

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

W.-L. Yang

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033798

Paper presented at the 2002 annual meeting of the National Council on Measurement in Education, New Orleans.

2 BEST COPY AVAILABLE

Abstract

Scores on the multiple-choice sections of alternate forms are equated through anchor-test equating for the Advanced Placement Program (AP) examinations. There is no linkage of free-response sections since different free-response items are given yearly. The free-response and multiple-choice sections are combined to produce a composite, however. To derive new-form cut scores on the composite score scale that are comparable to the old-form cut scores, therefore, the multiple-choice score of the AP exams are linked to the composite score through single-group linking design. This study investigates whether the multiple-choice to composite linking functions remain invariant over subgroups by region for two AP exams using three years of test data. The region groups are of interest because the AP program administers different free-response sections to different time-zone regions as a precaution for security reasons. The study focuses on two questions: (a) How invariant are cut scores across regions; (b) Does the small sample size for some regional groups present particular problems for assessing linking invariance? Both equipercentile and linear linking methods are applied. The equatability index proposed by Dorans and Holland (2000) is employed to evaluate the invariance of the linking functions, and the cross-classification approach is used to evaluate the invariance of the composite cut scores. Overall, the curvilinear linkings across regions seem to hold up reasonably well. Nevertheless, more exams need to be considered to decide whether the linkings in the small regions contain enough data to support the use of weak models such as the equipercentile scaling. The strong smoothing associated with the linear linking model, as expected, reduces the variability of linkages across regions. However, more work is needed to see if less strong models can be found for improving consistency over the unsmoothed linkings without incurring the bias apparent with the linear linkings.

1. Introduction and Overview

Test equating is a statistical process used to produce scores that are adequately comparable across interchangeable test forms. The goal of test equating is to produce fair and equitable measures. One of the most basic requirements of equating functions is that, to the extent possible, they should be population invariant (Dorans and Holland, 2000). That is, equating functions should not be strongly influenced by the population of examinees on which they are derived. One way to demonstrate that two tests are not equatable is to show that the "equating functions" used to link their scores are not invariant across different populations of examinees. No acceptable equating function can ever be completely population invariant, even in the best of circumstances. Instead, in the situations where equating is usually performed, the dependence of the equating function on the population used to compute it is small enough to be ignored.

Advanced Placement Program (AP) examinations are equated via an internal anchor-test design in which the linking items are restricted to be multiple-choice items, because the free-response sections of the tests are disclosed at the time of the administration. The free-response sections and the multiple-choice sections measure the construct domain somewhat differently. As a result, the linking of composite (multiple-choice and free-response combined) scores may violate one of the requirements of equating that helps insure population invariance, "strict content representativeness". In other words, equating through the multiple-choice section may or may not serve as an adequate substititue for equating the composite score.

Dorans, Holland, Thayer and Tateneni (2002) examined invariance of equatings for three AP exams across gender groups. Tateneni and Dorans (2002) looked at the invariance issues for ethnic groups as well as gender comparisons on one AP exam. The present study examines regional invariance of linking/scaling for two AP exams. Our motivation for studying regional invariance was practical. Since it is easy to memorize the questions in the free-response section, there is some concern that this section may be readily compromised across time zones. As a precaution, the AP program administers different free-response sections to different time-zone regions.

AP examinations are administered internationally, though primarily in the United States. The administrations span many time zones. Regional invariance of scaling is important, particularly when different free-response sections are given to different time-zone regions. The bulk of national testing occurs within the Eastern to the Pacific Time zones. Most international testing (mainly in Canada and South America) also occurs in these core time zones. Alaska and Hawaii fall outside that range, however. Candidates from Alaska and Hawaii take the exam after it is given to the candidates in the core time zones. In contrast, all testing in Asia and Europe occurs before the testing in the core time zones. For the purposes of studying regional invariance, we decided to partition the AP candidate group based on where the candidates are from and when they take the exams. We created the following three geographical clusters: the **c**ore time zones, mostly mainland U.S.; the **b**efore **c**ore time zones, mostly Europe and Asia, where testing occurs before the testing in the core time zones; and the **a**fter **c**ore time zones, mainly Hawaii and Alaska, where testing occurs after the testing in the core time zones.

Since the number of candidates in the non-core time zones is dwarfed by the number in the core zones, the amount of data available for determining composite score thresholds from multiple-choice score thresholds is limited. The present research focuses on two particular questions:

1. How invariant are thresholds on selected AP exams across regions?
2. Do the reduced regional samples present particular problems for assessing threshold invariance?

For the first question—*How invariant are thresholds on selected AP exams across regions*, we examined three years (1999-2000) of candidate performance data for two AP course subjects respectively:

1) A **large**-volume test (English Literature and Composition) that exhibited inconsistent regional differences between the multiple-choice and composite sections.

2) An **intermediate**-volume test (Microeconomics) that exhibited inconsistent regional differences between the multiple-choice and composite sections.

For the second question—*Do the reduced regional samples present particular problems for assessing threshold invariance*, we computed the Dorans and Holland (2002) equatability indices. We used the recently developed measures of subpopulation

invariance for linear and non-linear equating methods (Dorans & Holland, 2000) as the major means of assessing equatability of the two AP Subjects exams. This approach requires performing linkings in subgroups as well as in the full population. Classification consistency is employed to assess grade invariance. Both the equatability indices and the grade invariance measures are described in Holland (2002).

For the second question, we also calculated the cross-classification agreements in AP grades to contrast the outcomes of linkings between the multiple-choice scores and the composite scores, and between each regional group and the total group. For the same set of comparisons, we applied both the unsmoothed equipercentile method and the mean-sigma linear method for the linkings. In all cases, we placed the total-group multiple-choice grade thresholds to the composite score scale in each region, using composite score data for the regional group to establish the link between multiple-choice score and composite score.

This report has six sections including the introductory overview. Section 2 summarizes the data collection design. Section 3 briefly describes the equatability and grade invariance measures. In section 4, we describe the two AP exams we studied and the data used to examine the invariance of grade thresholds across regions. Section 5 presents the results of the equatability analyses. In the last section, we summarize analyses results, discuss their implications for the AP exams, and highlight the implications of our findings for future practice.

2. Linking AP Scores: Data Collection Design and Linking Procedures

AP score linking involves two kinds of linkage:

A. Linking of the multiple-choice sections through a non-equivalent group common item design;

B. Linking of the multiple-choice score and composite score, the sum of the weighted free-response and multiple-choice scores, via a single group design.

Note that there is no linkage of free-response sections. Instead, the free-response section is combined with the multiple-choice section to produce a composite. This composite is

linked to the multiple-choice section (B), which itself has been placed on the AP base scale via a common item equating (A).

In general, the AP score data are organized as follows:

	$COMP_o$	MC_o	EQ	MC_n	$COMP_n$
P_o	✓	✓	✓		
P_n			✓	✓	✓

where COMP, and MC represent the composite and multiple-choice raw scores respectively, while EQ represents the raw score for common items in the multiple-choice anchor test. The o and n represent the two administrations, old and new, and P_o and P_n represent the two administration populations. A "✓" indicates that data are available for the column variable in the population indicated by the row, and a blank space indicates that data are missing/impossible. The common item set EQ is the only test that provides a link between the two administrations.

Generally, the composite is computed by the following equation:

$$COMP = a*(MC) + SWEP = SWOP + SWEP,$$

where SWEP is the "sum of weighted essay part" of the test and a is the weight applied to the MC part to obtain the SWOP, the "sum of weighted objective part".

In this study, the multiple-choice linkage across years is held constant. The single group design linking the SWOP (a*MC) to the composite was repeated using data for the total group, core (**C**) group, before-core (**BC**) group, and after-core (**AC**) group respectively. Both equipercentile and linear scalings were obtained from this single-group design.

3. Measures of Population Invariance of Linking Functions

Dorans and Holland (2000) denoted the two tests to be linked by X and Y, and let the appropriate observed scores from these two tests be x and y respectively (usually raw scores but sometimes scaled scores, such as in some of our examples from section 4). In

addition, all of their definitions were stated at the population level, and they denoted a population of examinees by $\boldsymbol{P}$, with subpopulations of $\boldsymbol{P}$ denoted by subscripts, such as $\boldsymbol{P}_j$. For our purposes, the set of subpopulations, $\{\boldsymbol{P}_j : j = 1, 2, \ldots\}$, will always *partition* $\boldsymbol{P}$ into a set of *mutually exclusive* and *exhaustive* subpopulations, i.e., core group, before-core group, and after-core group.

Suppose we compute a linking function that links the scores on Y to those on X using the data from $\boldsymbol{P}$. Denote this linking function by $x = e_{\boldsymbol{P}}(y)$. The function $e_{\boldsymbol{P}}(\cdot)$ links scores on Y to equivalent scores on X. Dorans and Holland (2000) defined several measures of population dependence of linking functions. The paper by Holland (2002) discusses relevant measures to the present study.

In many tests, such as the SAT I, score equating functions are the focus of invariance studies (Dorans & Feigenbaum, 1994). With AP, the score linking function is an intermediate step on the way to the construction of the AP grade assignment rule. AP grades are reported on a 5-point scale, and the grades are obtained by applying cut scores to the new-form composite scale. The equatings and scalings for AP exams are used to convert scores to grades. Therefore, in addition to assessing the invariance of score linking functions, which are in the metric of the new-form composite score, we also need to study whether scalings based on different subpopulations produce different grade assignment rules. More importantly, we need to know whether the assignment rules based on subpopulation scalings result in grade assignments that differ from each other and from the grade assignments based on the total population scaling. Tateneni and Dorans (2002) provide an approach for assessing the consistency of classifications based on scalings for different subpopulations. Their approach is also described in Holland (2002).

4. AP Examinations Investigated

The AP exams we studied are the English Literature and Composition exam and the Microeconomics exam. The former is a large-volume exam while the latter is not. We chose English Literature exam because it is from the language domain. We expected regional differences, especially between the before-core region (Europe and Asia) and the

total group, in their relative experience with multiple-choice testing. English Literature exam was also selected because it gives more weight to the free-response section (55%) than to the multiple-choice section (45%).

The Microeconomics exam, an exam for a social science subject, was selected largely because it contains sparse data in the non-core groups. Another reason is that the multiple-choice score of the exam constitutes 2/3 of the composite score.

In a preliminary study comparing AP candidate's performance on the multiple-choice section and the composite, we found persistent evidence of possible time-zone dependency for the English Literature exam. There was emerging evidence of possible time-zone dependency for the Microeconomics exam.

5. Results of Equatability Analyses

5.1 Analyses Outcomes for the English Literature Exam

5.1.1 Equatability Results

The performance of the total group, the core, before-core and after-core groups on the composite score and SWOP score for the AP English Literature exam administered in 1999 are summarized in the top portion of Table 1. Table 1 includes the number of examinees, the proportion that each group contributes to the total group, the means and standard deviations of the composite and SWOP scores, and the Root-Expected-Mean-Square-Difference (REMSD) measures of equatability defined in Holland (2002).

Table 1: Summary Statistics for English Literature Exam Administered in 1999

Group	Total		Core		Before Core		After Core	
N	174,360		171,923		1,392		1,045	
Proportion (w_j)	1.000		0.986		0.008		0.006	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Composite	83.21	21.12	83.17	21.14	85.70	20.28	86.03	18.95
MC (SWOP)	39.63	12.45	39.62	12.45	40.47	12.29	41.18	11.65
	Overall		**Contribution by Region**					
REMSD Eq%	0.009		0.001		0.071		0.074	
REMSD μ/σ -1	0.006		0.001		0.059		0.043	
REMSD μ/σ -2	0.059		0.058		0.082		0.071	

The last three rows of Table 1 contain the REMSD statistics for equipercentile and linear linkings. The REMSD statistic needs further explication. REMSD is a weighted average of differences between subpopulation linking functions and the total group linking function. It is a double-weighted average. First, at each score level y, the difference between each subpopulation linking function and the total group linking function is squared. These squared differences are then averaged over subpopulations weighted by the relative size of each subpopulation (second row of Table 1). Then these weighted sums of squared differences are averaged across score levels weighted by the relative number of candidates in the total population at each score level. Finally, taking the square root of that weighted average and dividing the result by the standard deviation of the composite score in the total population gives us a measure of overall equatability in the metric of the standard deviation of the composite score.

How big a REMSD is big? In other words, when is REMSD large enough to evoke concern about the equatability of two exams? Dorans and Feigenbaum (1994) used the notion of *score differences that matter* (DTM) to answer this question in the context of linking the SAT I to the old SAT. On the SAT scales, scores are reported in 10-point units (200, 210, 220, ... , 780, 790, 800). At a given raw score point, linkings that produce linked scores that were within 5 points of each other were treated as close enough to be ignored because they were less than half of a reported score unit. This practice of ignoring differences that were less than half a score reporting unit has been used for SAT equating decisions for at least 15 years. We have adapted this practice to our present study.

There are two score metrics of interest with AP: the composite score metric and the AP grade scale. The unit of the composite score scale is one point. Likewise the unit of the grade scale is one grade. The range of the composite scale is from 0 to the sum of the maximum scores on SWEP and SWOP, and it varies from exam to exam. AP grades range from 1 to 5 for all exams. A unit on the AP grade scale is 20% of the range. A unit of one point on the composite score scale for the Microeconomics exam is 1.1% of the scale range, and for the English Literature exam it is 0.7%.

As the composite score approaches and crosses a grade threshold, a difference of one composite score means a change in AP grade. Hence, half a composite score

difference defines a DTM (*difference that matters*) on the composite score when that score is at a threshold. Elsewhere, large differences may not even matter. On the AP grade scale, a grade change is the DTM.

To obtain the standard score equivalent of a composite score of 0.5, we can simply divide 0.5 by the standard deviation. For the 1999 English Literature exam, this is 0.5 divided by 21.12, or 0.02. The overall REMSD for the equipercentile method is 0.009, well below the DTM. Since the overall REMSD measure is heavily influenced by the preponderance of core-group member (98.6%) in the total group, we need to examine the contribution of each subpopulation to the overall REMSD[1]. As expected, examination of these numbers, which appear alongside the overall index in Table 1, reveals that the core group and the total group linking are very close. In contrast, the other groups have individual REMSDs of about 0.070.

Root mean squared deviations are sensitive to the number of observations that they are based on. Equipercentile linking functions are also very sensitive to the number of observations. When the number of observations is small, these two sensitivities combine to produce large values of REMSD. One way to combat this REMSD inflation is to smooth the frequency distributions. Linear linking, which matches the first two moments of the distributions for the scores to be linked, is in essence a strong method of smoothing. As shown in the next to last row of Table 1, linear linking of the composite and SWOP scores only leads to a slight reduction in the overall REMSD, from 0.009 to 0.006. While there is no improvement in the large core group, there are noticeable reductions in the individual REMSDs in the before-core and after-core groups. This result is consistent with the expectation that linear linkings are more robust to small samples than curvilinear linkings.

The last row of Table 1 answers a more interesting question of how well the linear linkings in the regions approximate the total group curvilinear linking between SWOP and composite scores. As in the previous row, the linkings for the core, before-core and after-core groups are linear; while like the row above the previous row, the linking for the

[1] To compute the overall REMSD, square the individual regional REMSD, weight them proportionally by the relative size of each region, sum and take the square root to obtain a weighted average in the metric of a standardized composite score.

total group is curvilinear. The overall equatability index is 0.059. Contrast this number with the 0.009 for the curvilinear linkings. It indicates that, on average, across the three subgroups the linear equatings do not approximate the total curvilinear equating well. Closer examination of the individual REMSDs by region, however, reveals that the poorer overall fit is primarily a function of the large weight given to the poorer fit in the core group, which makes up 98.6% of the total group. In the two smaller regional groups, the difference between the agreement of the regional linear with the total group curvilinear and the agreement of the regional curvilinear with the total group curvilinear is much smaller. In fact, there is a slight improvement in the after-core group (from 0.074 for curvilinear to 0.071 for linear linking). This closer look underscores the pitfalls associated with relying on measures of global agreement such as the overall REMSD when one group is much larger than the other groups.

The performance of the total group, core, before-core and after-core groups on the composite score and SWOP score for the English Literature exam administered in 2000 are summarized in the top portion of Table 2. Table 2 includes the number of examinees, the proportion that each group contributes to the total group, the means and standard deviations of the composite and SWOP scores, and the REMSD measures of equatability. The last three rows of Table 2 contain REMSD statistics for equipercentile and linear linkings.

Table 2: Summary Statistics for English Literature Exam Administered in 2000

Group	**Total**		**Core**		**Before Core**		**After Core**	
N	98,877		97,522		736		619	
Proportion (w_j)	1.000		0.986		0.007		0.006	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Composite	80.06	23.00	80.03	23.01	82.13	23.51	82.40	20.55
MC (SWOP)	36.32	13.14	36.31	13.14	36.32	13.69	36.84	12.28
	Overall		**Contribution by Region**					
REMSD Eq%	0.011		0.001		0.096		0.097	
REMSD μ / σ -1	0.010		0.001		0.091		0.077	
REMSD μ / σ -2	0.051		0.050		0.104		0.091	

The standard score equivalent of a 0.5 composite score, the DTM, for the 2000 English Literature exam is 0.02, which is 0.5 divided by the standard deviation of 23.0. The overall REMSD for the equipercentile method is 0.011. Although the overall equatability measure is smaller than the DTM, it once again reflects the influence of the large core group on the overall measure of equatability. When we examined the contribution of each subpopulation to the overall REMSD, we found that, as expected, the core group and the total group linking are very close (98.6% of the total group comes from the core group). In contrast, the other groups have individual REMSDs of about 0.100. As shown in the next to last row of Table 2, linear linking of the composite and SWOP scores leads to only a slight reduction in the overall REMSD from 0.011 to 0.010. In the two smaller subgroups, however, there is small improvement.

The last row of Table 2 reveals again that the linear linking in the core group does not agree with the total-group curvilinear linking as well as the core-group curvilinear linking does. The linear linking in the before-core group almost agrees as well with the total-group curvilinear linking as the before-core group curvilinear linking does. In the after-core group, the linear linking actually agrees with the total-group curvilinear linking slightly better than does the after-core curvilinear linking.

The performance of the total group, core, before-core and after-core groups on the composite score and SWOP score for the AP English Literature exam administered in 2001 are summarized in the top portion of Table 3. Table 3 includes the number of examinees, the proportion that each group contributes to the total group, the means and standard deviations of the composite and SWOP scores, and the REMSD measure of equatability. The last three rows of Table 3 contain REMSD statistics for equipercentile and linear linkings.

Table 3: Summary Statistics for English Literature Exam Administered in 2001

Group	Total		Core		Before Core		After Core	
N	99,271		98,081		738		452	
Proportion (w_j)	1.000		0.988		0.007		0.005	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Composite	83.58	22.63	83.54	22.64	87.36	22.22	86.63	19.77
MC (SWOP)	39.13	13.23	39.11	13.23	40.28	13.48	40.13	12.50
	Overall		Contribution by Region					
REMSD Eq%	0.011		0.001		0.101		0.109	
REMSD μ / σ -1	0.010		0.001		0.090		0.099	
REMSD μ / σ -2	0.062		0.061		0.108		0.115	

The standard score equivalent of a 0.5 composite score, the DTM, for the 2001 English Literature exam is 0.02, which is simply 0.5 divided by the standard deviation of 22.6. The overall REMSD for the equipercentile method is 0.011. Despite that the overall equatability measure is smaller than the DTM, it is again heavily under the influence of the large core group. We examined the contribution of each subpopulation to the overall REMSD and found that, as expected, the core group and the total group linking are very close (98.8% of the total group comes from the core group). In contrast, the other groups have individual REMSDs of about 0.10. As shown in the next to last row of Table 3, linear linking of the composite and SWOP scores merely leads to a slight reduction in the overall REMSD, from 0.011 to 0.010.

The last row of Table 3 contains a fairly large overall REMSD index of 0.062, relative to the 0.011 of the REMSD for equipercentile linking. Once again, it is the large core group that drives this index. In the smaller groups, in terms of the agreement with the total-group curvilinear linking, the linear linking performs almost as well as the curvilinear linking.

5.1.2 Grade Consistency Results

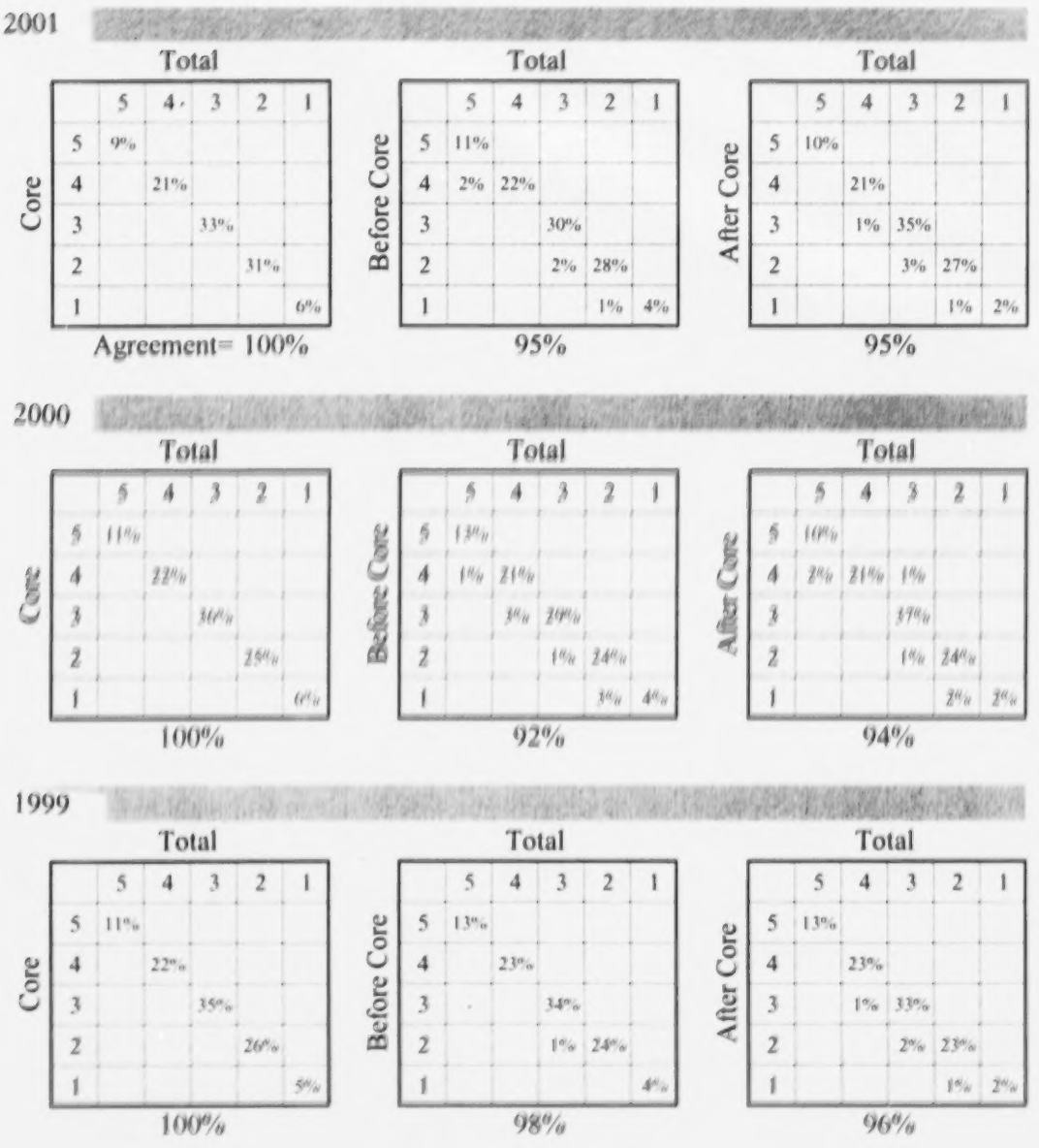
While invariance of multiple-choice to composite linkings is important because it undergirds consistency of grade assignments, it is the grade consistency that matters the most to AP. The consistency of the mapping from the composite score to AP grades

thresholds is summarized in a series of 5×5 cross-classification displays. In each set of the displays, there are three rows (by year) and three columns (by subgroup) of the 5×5 classification tables. Each of the 5×5 table contains the classification agreements and/or disagreements between the grade assignments based on the total-group linking (columns) and the grade assignments based on the subgroup linking (rows).

The three rows of the 5×5 tables in each set of display correspond to the years 2001, 2000, and 1999. The three columns of the tables in the display correspond to the core time-zone group, the before-core group, and the after-core group. The cross-classifications within each of the nine tables in the display are classifications based on total-group grade thresholds versus the subgroup-specific thresholds.

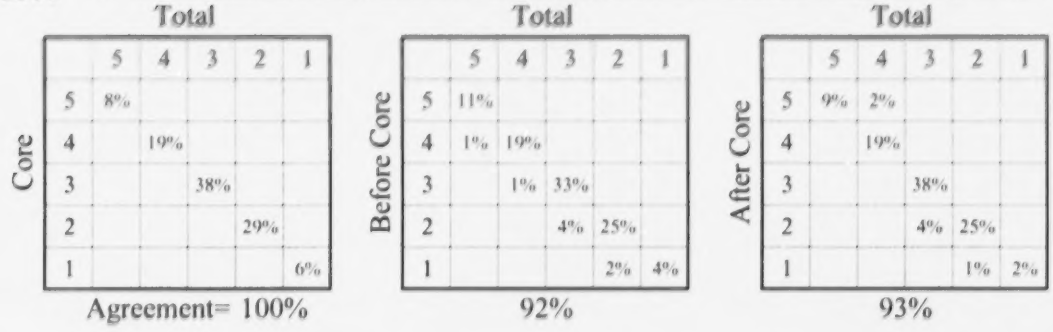
In every table that follows the agreement is perfect for the core-group members, which make up the vast majority of all candidates. For the large core group, overall the agreement is excellent.

AP English Literature & Composition Exam (Equi%tile Scaling)

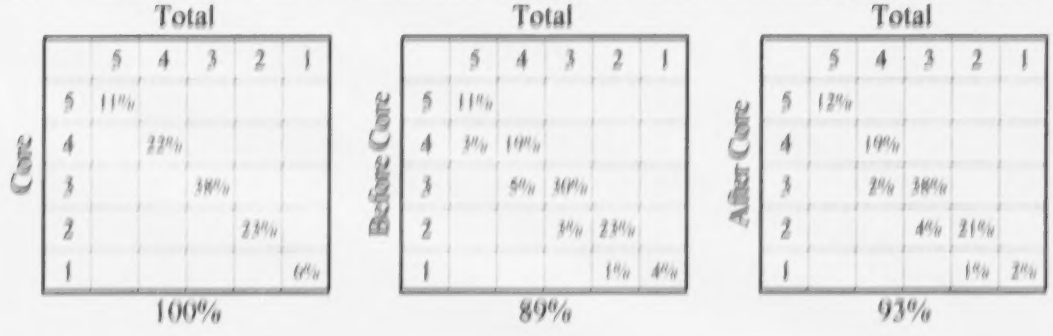


The above display summarizes the cross-classification results based on the equipercentile linkings. In general, the results are quite good. Even in the worst case, the before-core group in 2000, the overall agreement (sum of the main diagonal elements) is 92%.

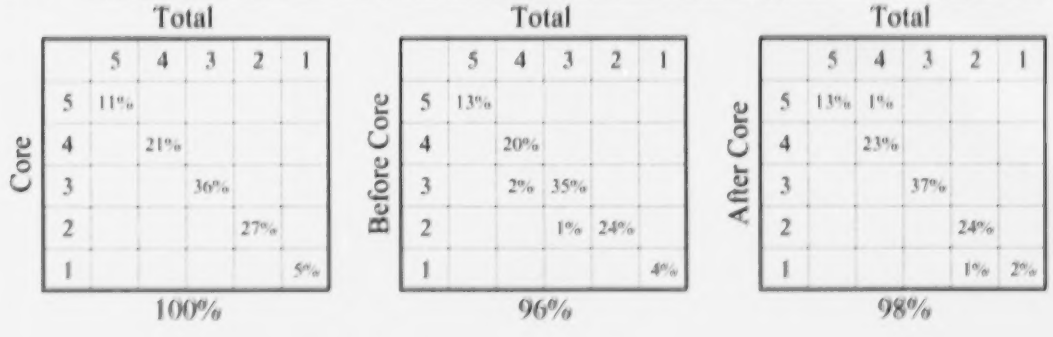
2001



2000



1999



The cross-classification outcomes based on the linear linkings appear above. In general, the results are almost as good as those based on the curvilinear linkings. When the linear linkings are used to approximate the total-group curvilinear linking, however, the results are much poorer. In general, it seems that the strong smoothing afforded by the linear linking does not improve grade assignment consistency.

5.2 Analyses Outcomes for the Microeconomics Exam

5.2.1 Equatability Results

The performance of the total group, the core, before-core and after-core groups on the composite score and SWOP score for the AP Microeconomics exam administered in 1999 are summarized in the top portion of Table 4. The last three rows of Table 4 contain REMSD statistics for equipercentile and linear linkings.

Table 4: Summary Statistics for Microeconomics Exam Administered in 1999

Group	Total		Core		Before Core		After Core	
N	14,797		14,133		442		222	
Proportion (w_j)	1.000		0.955		0.030		0.015	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Composite	47.40	18.26	47.15	18.24	51.46	18.17	55.25	16.55
MC (SWOP)	33.96	12.85	33.80	12.84	36.59	12.98	39.23	11.33
	Overall		Contribution by Region					
REMSD Eq%	0.013		0.002		0.044		0.089	
REMSD μ / σ -1	0.006		0.001		0.026		0.029	
REMSD μ / σ -2	0.031		0.030		0.040		0.042	

To obtain the standard score equivalent of a 0.5 composite score, the DTM, we divided 0.5 by the standard deviation of 18.3. For the 1999 Microeconomics exam, the DTM is 0.03. The overall REMSD for the equipercentile linking is 0.013, about 1/3 of this stringent DTM. However, since the overall REMSD measure is so heavily influenced by the preponderance of core-group member (95.5%) in the total group, we need to examine the contribution of individual subpopulation to the overall REMSD. Examination of these numbers reveals that, as expected, the core group and the total group linkings are very close. In contrast, the other groups have individual REMSDs of 0.04 and 0.09.

As shown in the next to last row of Table 4, linear linking of the composite and SWOP scores leads to a reduction in REMSD from 0.013 to 0.006. While there is only slight improvement in the large core group, there are noticeable reductions in the

individual REMSD measures in the before-core and after-core groups. This result is consistent with the expectation that linear linkings are more robust to small samples than curvilinear linkings.

The overall equatability index in the last row of Table 4 is 0.031. It seems that, on average, across the three subgroups the linear equatings reasonably approximate the total curvilinear equating well. Closer examination of the individual REMSD indices by region, however, reveals that there is a noticeable improvement over the curvilinear linking in the after-core group (from 0.089 to 0.042).

The performance of the total group and various subgroups on the composite score and SWOP score for the Microeconomics exam administered in 2000 are summarized in the top portion of Table 5. The last three rows of Table 5 contain REMSD statistics for equipercentile and linear linkings.

***Table 5: Summary Statistics for Microeconomics Exam* Administered in 2000**

Group	Total		Core		Before Core		After Core	
N	17,033		16,368		488		177	
Proportion (w_j)	1.000		0.961		0.029		0.010	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Composite	46.13	18.95	45.89	18.98	53.05	17.02	48.86	18.14
MC (SWOP)	32.77	13.01	32.62	13.03	37.14	11.81	34.71	11.93
	Overall		Contribution by Region					
REMSD Eq%	0.013		0.001		0.066		0.070	
REMSD μ / σ -1	0.008		0.001		0.035		0.045	
REMSD μ / σ -2	0.028		0.027		0.044		0.053	

We obtained the standard score equivalent of a 0.5 composite score, the DTM, by dividing 0.5 by the standard deviation of 19. For the 2000 Microeconomics exam, the DTM is 0.03. The overall REMSD for the equipercentile method is 0.013. Although the overall equatability index is smaller than the DTM, the overall measure is once again under the influence of the large core group. When we examined the contribution of individual subpopulation to the overall REMSD, we found that, as expected, the core group and the total group linking are very close (96.1% of the total group comes from the core group). In contrast, the other groups have individual REMSDs of about 0.07. As

shown in the next to last row of Table 5, linear linking of the composite and SWOP scores leads to a reduction in the overall REMSD, from 0.013 to 0.008. In addition, in the two smaller subgroups, there are noticeable improvements.

The last row of Table 5 reveals that the linear linking in the core group does not agree with the total-group curvilinear linking as well as the core-group curvilinear linking does. In both the before-core and after-core groups, the linear linking is closer to the total group curvilinear linking than are the equipercentile linkings.

The performance of the total group and various subgroups on the composite score and SWOP score for the AP Microeconomics exam administered in 2001 are summarized in the top portion of Table 6. The last three rows of Table 6 contain REMSD statistics for equipercentile and linear linkings.

Table 6: Summary Statistics for Microeconomics Exam Administered in 2001

Group	**Total**		**Core**		**Before Core**		**After Core**	
N	18,030		17,427		372		231	
Proportion (w_j)	1.000		0.967		0.021		0.013	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Composite	48.29	19.60	48.05	19.61	54.28	17.53	56.92	18.71
MC (SWOP)	34.65	12.91	34.53	12.93	37.38	11.33	39.04	11.97
	Overall		**Contribution by Region**					
REMSD Eq%	0.023		0.004		0.111		0.137	
REMSD μ/σ -1	0.017		0.004		0.091		0.094	
REMSD μ/σ -2	0.059		0.056		0.108		0.111	

The standard score equivalent of a 0.5 composite score, the DTM, is simply 0.5 divided by the standard deviation of 19.6. For the 2001 Microeconomics exam, this is 0.03. The overall REMSD for the equipercentile method is 0.023. We examined the contribution of each subpopulation to the overall REMSD and found that, as expected, the core group and the total group linkings are very close (96.7% of the total group comes from the core group). In contrast, the smaller groups have individual REMSDs of 0.11 and 0.14. As indicated in the next to last row of Table 6, linear linking of the composite and SWOP scores leads to a slight reduction in the overall REMSD, from 0.023 to 0.017.

While there is no improvement in the core group, there are improvements in the individual REMSDs for the two smaller subgroups.

The last row of Table 6 contains a fairly large overall REMSD index of 0.059, relative to the 0.023 for the equipercentile linking. Once again, it is the large core group that drives this index. In the smaller groups, in terms of the agreement with the curvilinear total-group linking, the linear linking performs as well as, if not better, than the curvilinear linking.

5.2.2 Grade Consistency Results

In every table that follows, for the core group, the agreement between the grade assignments based on the total-group linking (columns) and the grade assignments based on the subgroup linking (rows) is perfect. Note that the core group makes up the vast majority of the total group.

AP Microeconomics Examination (Equi%tile Scaling)

2001

2001

Total

Core	5	4	3	2	1
5	14%				
4		27%			
3			22%		
2				21%	
1					16%

Agreement= 100%

Total

Before Core	5	4	3	2	1
5	17%				
4	5%	26%			
3		8%	17%		
2			1%	18%	
1				2%	8%

85%

Total

After Core	5	4	3	2	1
5	27%				
4	3%	25%			
3		3%	17%		
2			3%	14%	
1				2%	7%

90%

2000

2000

Total

Core	5	4	3	2	1
5	10%				
4		27%			
3			23%		
2				21%	
1					18%

100%

Total

Before Core	5	4	3	2	1
5	15%				
4		35%			
3		2%	20%		
2			1%	19%	
1					8%

97%

Total

After Core	5	4	3	2	1
5	13%				
4	3%	24%			
3			26%		
2				21%	
1				2%	11%

94%

1999

1999

Total

Core	5	4	3	2	1
5	12%				
4		27%			
3			23%		
2				23%	
1					14%

100%

Total

Before Core	5	4	3	2	1
5	18%				
4		32%			
3			19%		
2			2%	16%	
1					12%

97%

Total

After Core	5	4	3	2	1
5	21%				
4	3%	29%			
3		1%	24%	1%	
2				17%	
1					4%

95%

The above display summarizes the cross-classification outcomes based on the equipercentile linkings for the Microeconomics exam. Overall, the classifications are quite consistent except for 2001. For 2001, the agreement is only 89% in the after-core group and only 84% in the before-core group.

BEST COPY AVAILABLE

AP Microeconomics Examination (Linear Scaling)

2001

2001

Total

Core	5	4	3	2	1
5	16%				
4		24%			
3			22%		
2				23%	
1					16%

Agreement= 100%

Total

Before Core	5	4	3	2	1
5	19%				
4	5%	27%			
3		2%	16%		
2			3%	19%	
1				2%	8%

88%

Total

After Core	5	4	3	2	1
5	26%				
4	5%	23%			
3		3%	14%		
2			5%	17%	
1					7%

87%

2000

2000

Total

Core	5	4	3	2	1
5	10%				
4		25%			
3			25%		
2				21%	
1					18%

100%

Total

Before Core	5	4	3	2	1
5	15%				
4		35%			
3			22%		
2			1%	18%	
1				1%	8%

98%

Total

After Core	5	4	3	2	1
5	15%				
4	1%	22%			
3			28%		
2				23%	1%
1					10%

98%

1999

1999

Total

Core	5	4	3	2	1
5	12%				
4		27%			
3			23%		
2				23%	
1					14%

100%

Total

Before Core	5	4	3	2	1
5	18%				
4		30%			
3		3%	19%		
2			2%	16%	
1				1%	12%

94%

Total

After Core	5	4	3	2	1
5	24%				
4		29%			
3		1%	24%		
2				18%	1%
1					3%

98%

The cross-classification outcomes based on the linear linkings for the Microeconomics exam appear above. In general, the results are similar to, if not better than, those seen for the curvilinear linkings. When the linear linkings are used to

approximate the total-group curvilinear linking, however, the classification results are worse than the curvilinear linkings. For the Microeconomics exam, for which the volume is much smaller than the other exam studied, the strong smoothing of linear linking may not improve the classification consistency.

6. Summary

Curvilinear linkings across regions seem to hold up reasonably well for all the English Literature exams, and for the 1999 and 2000 Microeconomics exams. Nevertheless, more exams need to be considered before we can conclude that the linkings associated with the before-core and after-core groups contain enough data to support the use of weak models such as equipercentile scaling for the linking of the multiple-choice score to the composite score.

The strong smoothing associated with the linear linking model, as expected, reduces the variability of linkages across regions. However, this reduction comes at the expense of increased bias. The smoothing of linear linking is the equivalent to matching the first two moments of the score distributions in Holland and Thayer's (1998) log-linear smoothing model. More work needs to be done to see if less strong models can be found that will improve consistency over the unsmoothed linkings without incurring the bias apparent with the linear linkings.

The existence of a dominant subpopulation such as the core group underscores the need to supplement the overall Dorans and Holland measure with subpopulation-specific REMSD measures. Otherwise, inconsistency in some subpopulations may be overlooked because of their small sizes. Nevertheless, the REMSD measure itself is biased in small samples. Therefore, care needs to be taken to avoid over interpretation of these subpopulation measures in small samples.

References

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright, *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM-94-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2002, April). *Invariance of score linking across gender groups for three Advanced Placement Program Exams.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Holland, P. W. (2002, April). *Overview of population invariance of score linking*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Holland, P. W., & Thayer, D. T. (1988). *Univariate and bivariate loglinear models for discrete test score distributions.* (Technical Report No. 98-1). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (RR-89-7). Princeton, NJ: Educational Testing Service.

Tateneni, K., & Dorans, N. J. (2002, April). *Invariance of linkages for free response, multiple-choice and composite scores on Advanced Placement Program examinations.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

TM033798

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Sample selection effect on AP multiple-choice score to composite score scaling	
Author(s): Wen-Ling Yang, Neil J. Dorans, Krishna Tateneni	
Corporate Source: Educational Testing Service	Publication Date: April 3, 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ↑ ☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑ ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑ ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Wen-Ling Yang	Printed Name/Position/Title: Wen-Ling Yang, Measurement Statistician	
Organization/Address: Educational Testing Service, Rosedale Road, MS 11-L, Princeton, NJ 08541	Telephone: 609-683-2688	FAX: 609-6832400
	E-Mail Address: Wyang@ets.org	Date: 4/1/2002

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)